BlackboxNLP2020

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-952148-86-6

Introduction

BlackboxNLP is the workshop on analyzing and interpreting neural networks for NLP.

In the last few years, neural networks have rapidly become a central component in NLP systems. The improvement in accuracy and performance brought by the introduction of neural networks has typically come at the cost of our understanding of the system: How do we assess what the representations and computations are that the network learns? The goal of this workshop is to bring together people who are attempting to peek inside the neural network black box, taking inspiration from machine learning, psychology, linguistics, and neuroscience.

In this third edition of the workshop, hosted by the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP), we accepted 31 archival papers and 9 extended abstracts. The workshop also provided a platform for authors of EMNLP-Findings papers to present their work as a poster at the workshop. Lastly, for the first time, BlackboxNLP included a shared interpretation mission. One paper submitted to this mission has a *demo* presentation, of the interpretability library `diagnnose`, and is included as the last paper in these proceedings (submission number 70).

BlackboxNLP would not have been possible without the dedication of its program committee. We would like to thank them for their invaluable effort in providing timely and high-quality reviews on a short notice. We are also grateful to our invited speakers for contributing to our program.

Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter and Hassan Sajjad

Organizers:

Afra Alishahi, Tilburg University
Yonatan Belinkov, Technion - Israel Institute of Technology
Grzegorz Chrupała, Tilburg University
Dieuwke Hupkes, University of Amsterdam
Yuval Pinter, Georgia Institute of Technology
Hassan Sajjad, Qatar Computing Research Institute

Program Committee:

Samira Abnar, University of Amsterdam
Željko Agić, Unity Technologies Copenhagen
Antonios Anastasopoulos
Leila Arras - Fraunhofer Heinrich Hertz Institute
Jasmijn Bastings - Google
Lisa Beinborn - University of Amsterdam
Laurent Besacier - Laboratoire d'Informatique de Grenoble
Stergios Chatzikyriakidis - University of Gotheburg
Barry Devereux - Queen's University
Ewan Dunbar - Université Paris Diderot
Allyson Ettinger - University of Chicago
Antske Fokkens - Vrije Universiteit Amsterdam
Robert Frank - Yale University
Alexander Fraser - LMU Munich
Richard Futrell - University of California, Irvine
Sebastian Gehrmann - Harvard University
David Harwath - MIT
John Hewitt - Stanford University
Cassandra Jacobs - University of Wisconsin
Yair Lakretz - NeuroSpin
Shalom Lappin - University of Gothenburg
Miryam de Lhoneux - Uppsala University
Tal Linzen - Johns Hopkins University
Nelson F. Liu - University of Washington
Pranava Madhyastha - Imperial College London
Arya McCarthy - Johns Hopkins University
Paola Merlo - University of Geneva
Raymond Mooney - UT AUstin Joakim Nivre - Uppsala University
Sebastian Padó - University of Stuttgart
Ellie Pavlick - Brown University
Rudolf Rosa - Charles University
Carolyn Rose - CMU Sebastian Ruder - DeepMind
Wojciech Samek - Fraunhofer Heinrich Hertz Institute
Naomi Saphra - University of Edinburgh
Sabine Schulte im Walde - University of Stuttgart
Rico Sennrich - University of Zurich
Pia Sommerauer - Vrije Universiteit Amsterdam

Ivan Titov - University of Edinburgh
Francesca Toni - Imperial College London
Reut Tsarfaty - Open University
Sarah Wiegreffe - Georgia Tech
Adina Williams - New York University
Diyi Yang - Georgia Tech
Fabio Massimo Zanzotto - University of Rome

Invited Speakers:

Idan Blank, UCLA
Roger Levy, MIT
Anna Rogers, University of Copenhagen

Table of Contents

Conference Program

The programme of BlackboxNLP2020 consists of three keynote presentations, six selected oral presentations, one demo paper and two poster sessions. Due to the virtual nature of the conference, these activities are distributed over three blocks, such that every activity occurs twice and is accessible for any time zone. The full programme of the workshop can be found at https://blackboxnlp.github.io.

The three keynotes at BlackboxNLP2020 are:

Roger Levy, MIT
Evaluating and calibrating neural language models for human-like language processing

Anna Rogers, University of Copenhagen
When BERT plays the lottery, all tickets are winning!

Idan Blank, UCLA
Understanding NLP's blackbox with the brain's blackbox and vice versa

The six papers selected for oral presentation are:

What Happens To BERT Embeddings During Fine-tuning?
Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick and Ian Tenney

Dissecting Lottery Ticket Transformers: Structural and Behavioral Study of Sparse Neural Machine Translation
Rajiv Movva and Jason Zhao

Evaluating Attribution Methods using White-Box LSTMs
Yiding Hao

The EOS Decision and Length Extrapolation
Benjamin Newman, John Hewitt, Percy Liang and Christopher D. Manning

BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance
R. Thomas Mccoy, Junghyun Min and Tal Linzen

The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?
Jasmijn Bastings and Katja Filippova

The shared task paper selected to give a demo presentation is:

diagNNose: A Library for Neural Activation Analysis
Jaap Jumelet

All other papers in these proceedings, as well as the nine accepted abstracts, are presented at the poster sessions of the conference. Also a selection of related EMNLP-findings papers are present at the poster sessions.